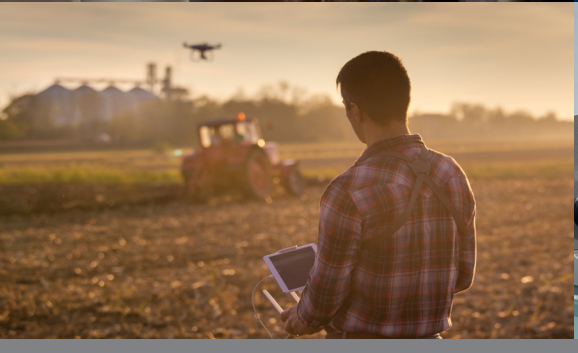
BURLEIGH DODDS SERIES IN AGRICULTURAL SCIENCE

Sampling and statistics in assessment of fresh produce

K. B. Walsh, Central Queensland University, Australia; and V. A. McGlone and M. Wohlers, The New Zealand Institute for Plant and Food Research Limited, New Zealand



burleigh dodds SCIENCE PUBLISHING

Sampling and statistics in assessment of fresh produce

K. B. Walsh, Central Queensland University, Australia; and V. A. McGlone and M. Wohlers, The New Zealand Institute for Plant and Food Research Limited, New Zealand

1 Introduction

The management of any supply chain requires accurate inventories of stock quantity and quality. This requirement is accentuated in fresh fruit and vegetable supply chains, which involve perishable produce with a narrow window on harvest timing and shelf life. For example, to reduce the potential for product loss, forward knowledge of both crop load and harvest timing is required before harvest to inform decision making on harvest resourcing in terms of requirements ranging from labour needs to purchase of packaging materials. Additional post-harvest information on product quality, shelf life and infestation is required to inform marketing and biosecurity assessments.

There have been great advances in technologies for the assessment of produce attributes in recent decades. For example, weight, colour and defect grading are now commonplace in packhouses, and technology for both in-line and in-field, non-invasive assessment of fruit soluble solids content (SSC, usually measured as °Brix) and dry matter is readily available. Various remote or proximal sensing technologies are also becoming available for the estimation of fruit load pre-harvest. However, as with all measurement scenarios, the step

http://dx.doi.org/10.19103/AS.2021.0097.09

Chapter taken from: Manning, L. (ed.), *Developing smart agri-food supply chains: Using technology to improve safety and quality*, Burleigh Dodds Science Publishing, Cambridge, UK, 2022, (ISBN: 978 1 78676 749 3; www.bdspublishing.com)

from data collection to interpretation requires consideration of the quality of the data in terms of both measurement accuracy and precision.

All measurements have the potential for error. Measurement error is low relative to the required specification in some cases, for example, weight assessment on a packing line, while in other cases the measurement error for the assessment of a given criterion is relatively high. The setting of a specification should, therefore, give consideration to the potential for such error, for example, a specification requiring that the level of an attribute be above a specific value might be shifted to a higher value to accommodate measurement error and to avoid inclusion of under-specification produce in the accepted class.

There is typically only one point in fruit and vegetable supply chains where all produce can be inspected as individual units – at the packhouse as the produce passes over a grading line (Fig. 1). Elsewhere in the chain from the orchard through to retail shelf, assessment is based on a sample only of the consignment. A decision must be taken on where to take samples and how many samples to take. Acquiring a sample that is representative of the population is not a trivial task, however, and sampling practice is often constrained by the logistics of undertaking these assessments. The issue is, of course, variability.

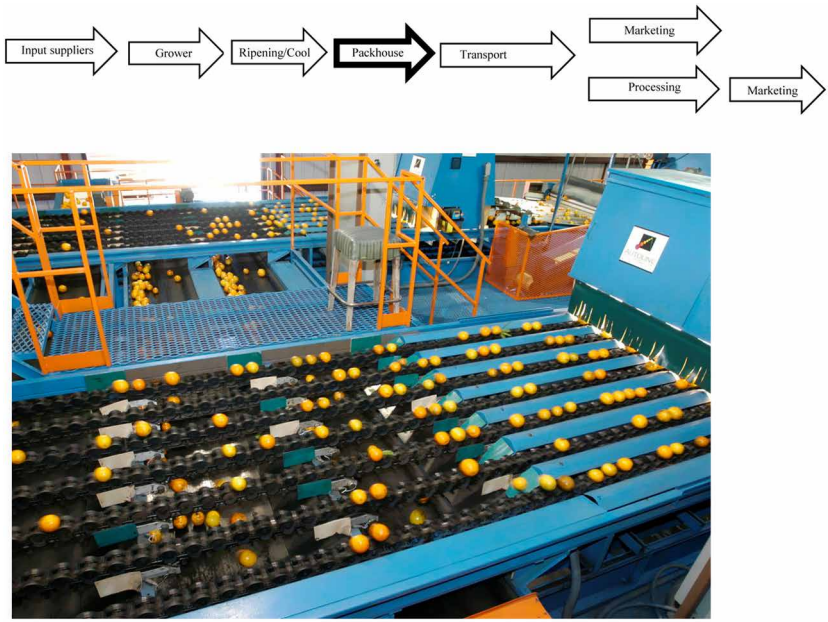


Figure 1 Typical chain involved in production, transport, storage and/or ripening and marketing of fresh produce (top panel) and an image of a grading line – the one point in the chain where all produce items can be inspected individually (bottom panel). (Image source: https://search.creativecommons.org/photos/aaa81a9d-38a4-4b72-9fa5-912b03c61eeb).

Published by Burleigh Dodds Science Publishing Limited, 2022.

Crudely, the more variability in the population, the greater the sample size required and the more attention that is required to the selection of samples to achieve a valid estimate of that variability and of the mean of the population.

Measurement and sampling issues are recognised in the scientific literature. For example, Follett and Hennessey (2007) note that 'incorporating sample size and confidence levels into host status testing protocols along with efficacy will lead to greater consistency by regulatory decision-makers in interpreting results and, therefore, to more technically sound decisions on host status'. However, while the issue is recognised, it is under-represented in scientific literature, relative to reports on measurement technology and its efficacy. For example, the text 'Fruit and vegetables, harvesting, handling and storage' (Thompson, 2003) gives comprehensive coverage to methods of the assessment of various fruit attributes but no coverage of population sampling.

A recent review concluded that 'further consideration of sampling statistics is warranted, with the aim of producing decision support aids to postharvest management, to assist in design of sampling regimes' (Walsh et al., 2020). We attempt to address this call in the current chapter. Relevant statistical tools are reviewed, with worked examples, and case studies are presented of the pre-harvest measurement of fruit load in an orchard, determining harvest timing and post-harvest assessment of eating quality or biosecurity assessment.

2 Positioning the industry problem

Consider the issue of pre-harvest crop load estimation. Fruit can be counted on a sample of plants, but for the estimate to be valid and representative, how many plants should be counted and how should those plants be selected? In Australia, the national citrus crop estimate involves surveys of the same sites assessed every year for the density of fruit on representative tree canopies across many orchard blocks, with fruit density counts based on counting fruit numbers in a 'frame' (Citrus Australia, 2003). However, the statistical rationale that forms the basis for such best-practice sampling regimes is not readily available.

Specifications on post-harvest quality attributes have been developed by international bodies, such as the Codex Alimentarius and the United Nations Economic Commission for Europe (UNECE), with these recommendations informing private organisation specifications and standards, for example, GlobalGAP or individual retailers. These specifications typically place quantitative limits on external product attributes, such as size, shape, colour, and surface defects, and on internal attributes, such as SSC or dry matter content, firmness and defects, such as internal browning in apple. However, whilst these specifications give detailed product criteria with targets and tolerances, they are generally weak with regard to the sampling protocol to be used in testing the degree of

compliance with the specification, especially in terms of sampling strategy and numbers (Fig. 2). For example, a typical specification for apple includes 'total minor defects (within allowance limit) to be <2 defects per item, total minor defects (outside allowance limit) must not exceed 10% of consignment, and total major defects must not exceed 2% of consignment, with combined total not to exceed 10%' (Freshmark, 2020). However, the specification is silent on the number of fruit to be sampled per consignment, that is, the sampling frequency. Likewise, the public documentation of those providing testing services, for example, QIMA (2020) or Farmsoft (2020), is also silent on the sampling strategy to be adopted.

Perhaps the best accessible documentation on sampling for fruit specifications is provided by the United States Department of Agriculture (USDA, 2021). The US Grade Standards are supplemented by documentation on sampling. This documentation notes that 'the importance of obtaining representative samples cannot be over emphasized. Accurate certification is possible only if the samples examined are truly representative of the entire lot or accessible portion. All portions of a lot or load shall receive the same attention in sampling regardless of the difficulty involved in reaching all layers or parts of a lot or load'.

In practice, the reality of sampling constrains this ideal and much can be left to operator interpretation. For example, the USDA instructions for banana are 'the sample size shall be a minimum of 50 count (50 individual bananas) for packages containing 50 or more specimens', with the number of samples involving a minimum of 1% of the lot. That instruction is clear; however, a caveat is added: 'It is the inspector's responsibility to examine additional representative samples when the quality or condition (of) samples is decidedly different to ensure an accurate description of the lot' (AMS, 2020a).

Figure 2 A consignment of fruit reaching a distribution centre. How should a representative sample of fruit be taken for quality control assessment?

Published by Burleigh Dodds Science Publishing Limited, 2022.

In another example, the instructions (AMS, 2020b) for the sampling of lemon consignments in bulk containers (e.g. trailers, bulk bins) are given as 'examine a minimum of 25 contiguous fruit per sample. When a sample tolerance is exceeded, the sample size must be at least doubled'. The caveat is added: 'due to potential variations in size, quality and condition, a specific number of samples per load or lot cannot be provided. It is the inspector's responsibility to examine a sufficient number of samples to ensure that a complete and accurate depiction of the load or lot is obtained'. Such instructions place considerable discretion and responsibility on the inspector.

Similar sampling issues arise in biosecurity assessments. For example, one regulator of horticultural produce biosecurity inspections requires a sample of either 600 units (pieces of fruit) or 2% (with a minimum of three packages) of the number of packages in each consignment (DAWR, 2021). These sampling criteria set a minimum number of items to sample, yet in practice the number of samples required to detect an infested sample at a desired confidence level will vary with incidence rate. These examples highlight the uncertainty in sampling strategies that can be found in both regulatory and market-based fresh produce sampling.

3 Sampling statistics

3.1 The problem is variation

Of course, if all units (i.e. individual fruit) in a lot were identical, only one sample (i.e. selection of fruit) would need to be assessed for a consignment description to be representative of all units in that lot. However, as the degree of variation for given criteria within the lot increases, more units must be assessed to provide a representative picture of the diversity of the consignment. Ideally all units in each lot would be assessed to achieve a true estimate of population parameters. Such measurement of the entire population is rarely possible, with some exceptions, for example, fruit passing sensors such as cameras or load cells on a packing line. To reduce measurement effort, a sample population of units is selected from the population and used to make inferences about the parameters of the whole population. The representativeness of this sample is determined by its size and the sampling strategy used to collect it.

The selection of sampling strategy and size is a vexed one, requiring preliminary information on the spatial and possibly temporal variation of the attribute of interest in the population for an informed decision on the choice of sampling strategy, and for an estimate of population variance to inform the choice of the number of samples used.

This section of the chapter gives background to the issues of 'where to sample' and 'how many to sample', to describe a population in terms of the

number of units, the average result and variance in the level of an attribute or the proportion of units meeting some specific criterion. These approaches and formulae are then implemented within several case studies.

3.2 Where to sample

To avoid bias in estimates, probability (random) sampling is required. The International Plant Sampling Standard ISPM 31of the International Plant Protection Convention (2020) describes the advantages and disadvantages of the common sampling strategies, including simple random sampling, systematic sampling, stratified sampling, sequential sampling, cluster sampling, fixed proportion sampling for statistically based sampling methods, convenience sampling, haphazard sampling and elective or targeted sampling for non-statistically based (non-probability) sampling. Each type of sampling affects the inferences that can be drawn from the data obtained from the analysis of attributes of the sampled population.

The following text considers four common approaches to sampling: simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Stratified sampling and clustered sampling involve multistage designs, while systematic sampling can be implemented as single or multistage designs.

1 *Simple random sampling*: In the simple random sample (SRS) scheme, units are randomly selected to be in the sample with equal probability, with each possible combination of units up to the sample size also equally likely to be selected. The standard statistical estimators of mean and variance may be used when using this sampling scheme.

   True random sampling is, however, often difficult to achieve in practice if sample selection is left to human choice. Operator bias exists in the 'random' selection of a sample, be that a tree in an orchard or a carton from a pallet. Best practice involves the assignment of numbers to units in the population and the use of a random number generator for the selection of samples.
2 *Systematic sampling*: Systematic sampling is easier to implement than SRS. In this sampling method, sample units are selected according to a random starting point but with a fixed, periodic interval. Population estimates are calculated as for SRS. However, this sampling scheme can result in a bias on estimates if there are spatially periodic effects in unit attribute levels. For example, sampling the carton from the same position in the top layer of every sampled pallet could result in a biased result if this position is not representative of the whole pallet. For example, if the supplier knows where the sampling from a given pallet will occur,

Published by Burleigh Dodds Science Publishing Limited, 2022.

they may place a higher quality product in this location. The sampling interval should, therefore, be considered carefully in implementing a systematic sampling strategy.

3 *Stratified sampling*: Stratified sampling requires the population to be divided into subgroups (or strata), which share a similar characteristic. This method can improve the accuracy and representativeness of the results by reducing sampling bias and can reduce the number of units needed. However, it can be difficult to choose the characteristic(s) to use in stratification.

4 *Clustered sampling*: Clustered sampling can be implemented with any of the above sampling strategies. This approach involves the use of subgroups ('clusters') of the population as the sampling unit, rather than individuals. Clusters are randomly selected for assessment. If the chosen clusters are not representative of the population, this method will result in sampling error.

Multistage designs involving sampling of units at different 'strata' within the population. Each stratum is assigned its own sampling strategy and probability. Consider a quality inspection of fruit in a shipping container containing pallets of fruit from two farms, with the sampling of five fruit per box of a box taken from each of three layers within pallets located at the front, middle and rear of the container, that is, a total of 90 fruit. This sampling strategy involves stratification on farm as the primary sampling unit (PSU), systematic sampling pallet and layer and random sampling of fruit within a box. In another example, consider the estimation of SSC of fruit in an orchard of 100 blocks of 1000 trees each, with orchard blocks defined by the management system (tree cultivar, age and training system). A sampling approach is imposed in which every third block of a given management system is randomly chosen, with assessment of five fruit randomly chosen from each side of every twentieth tree in every second row. In this approach, blocks are the PSU, with systematic sampling of blocks, rows and trees, stratified sampling of tree sides and random sampling of fruit on tree.

4 How many to sample: 'power' calculations in simple random sampling

In the following section, consideration is given to the question of estimating an appropriate sample size to collect for the determination of population mean to a known error and probability. Greater detail on sampling techniques and related equations can be found in Thompson (2012). Example applications include: determining the number of fruit that should be sampled to establish that the population average of an attribute is above a specified level and the number of trees that should be sampled in an orchard to establish the average

fruit load per tree. A key requirement for such estimates is a prior estimate of the population variance.

4.1 Does a lot meet specification?

Consider the determination of whether a population mean is greater than a specification. First the minimum detectable difference, *d*, should be defined. This value is set to be confident in concluding the population mean is above threshold from the sample if the population mean is at least *d* or greater above the threshold. Often *d* is based on domain knowledge and may be the smallest value to be biologically important. It should not be 0, however, as if the population mean is equivalent to the threshold, then the sample mean will be below the threshold half of the time. An estimate of the population mean can be used to inform the setting of *d*, with a smaller required sample size associated with the use of a larger *d* in association with a larger mean to specification difference.

A statistical test is then used to determine if there is evidence that the population mean is greater than the specification. The test will be undertaken to a given probability, called the power of the test. Often the power is set to 80%. This is equivalent to specifying the type 2 error at 20% ($\beta = 0.2$), which is the chance the test incorrectly fails (false negative) to detect that the population is above the standard. The chance of a type 1 (false-positive) error, that is, the probability that the test indicates that the population is above the standard when it is not, is defined by the setting of the significance level (α) of the test. This value is often set to 5% ($\alpha = 0.05$); that is, the user accepts a 1 in 20 chance of a false-positive result.

Given a simple random sampling, the minimum sample size needed to achieve a power of $(1-\beta)\times 100\%$ in detecting a difference *d* between the population mean and a standard using a t-test with $\alpha = 5\%$ is calculated as follows:

$$n = \frac{\sigma^2 \left(t_{\alpha,n-1} + t_{\beta,n-1}\right)^2}{d^2}, \tag{1}$$

where σ is the standard deviation, σ^2 is the variance, t_α and t_β are the values from the students' *t*-distribution with $n - 1$ degrees of freedom (*df*) such that the probability of being greater than t_α and t_β is α and β respectively. Note that *n* should always be rounded up to the nearest integer in giving the required sample size. There are many lookup tables available online for t_α and t_β, for example, https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm and https://stattrek.com/online-calculator/t-distribution.aspx. Additionally, various software packages include functions to calculate such values, for example, Microsoft Excel (T.INV) and R (qt).

Published by Burleigh Dodds Science Publishing Limited, 2022.

The t_α and t_β are a function of n; however, if n is large (>120 in SRS), the standard normal Z-score can be used:

$$n = \frac{\sigma^2 \left(Z_\alpha + Z_\beta\right)^2}{d^2}. \quad (2)$$

Z-score lookup tables are also available online, for example, https://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm, https://stattrek.com/online-calculator/normal.aspx, or through use of functions in Microsoft Excel (NORM.INV) and R (qnorm).

Consider a test to establish if a block of kiwifruit has an average dry matter (% FW; fresh weight) over a threshold of 16%. From a previous sample, the standard deviation was estimated at 2.0% and mean at 16.5%. It is desired to correctly accept blocks using a minimum detectable difference of 0.3% FW, that is, if block DM is >16.3% and to give the test 80% power (β = 0.2) to correctly determine that the population is above the threshold, while at the same time only having a 5% (α = 0.05) chance of wrongly concluding that the mean is above the threshold when it is not. To calculate the minimum sample size required, we can use eq. 2, with the input of the Z_α score for α = 0.05 and the Z_β score for β = 0.2:

$$n = \frac{2^2 \left(Z_{0.05} + Z_{0.2}\right)^2}{0.3^2} = \frac{4\left(1.645 + 0.8416\right)^2}{0.09} = 274.7$$

worked example for eq. (2).

Rounding up, we conclude 275 fruit are required.

If n were < 120, then eq. 1 should be used, starting with the solution from eq. 2 and increasing n until the $df = n - 1$. Consider if an α = 0.5 is used in the previous example. This is equivalent to a positive test if the sample mean is greater than the threshold irrespective of the magnitude. Using eq. 2:

$$n = \frac{2^2 \left(Z_{0.5} + Z_{0.2}\right)^2}{0.3^2} = \frac{4\left(0 + 0.8416\right)^2}{0.09} = 31.48$$

Rounding up, n is estimated to be 32. The value df in eq. 1 can be set as $n - 1$, i.e. 31, giving:

$$n = \frac{2^2 \left(t_{0.5,31} + t_{0.2,31}\right)^2}{0.3^2} = \frac{4\left(0 + 0.8534\right)^2}{0.09} = 32.37 = 33$$

Since the $df = 31 \neq 33 - 1$, we try again with n = 33, i.e. df = 32

$$n = \frac{2^2 \left(t_{0.5,32} + t_{0.2,32}\right)^2}{0.3^2} = \frac{4\left(0 + 0.8530\right)^2}{0.09} = 32.34 = 33$$

worked example for eq. (2).

Now the *df* and *n* match, so $n = 33$ is the final solution.

Sometimes it is useful to consider d as a percentage of a standard population mean. This is achieved by division by population mean, μ:

$$n = \frac{CV^2\left(Z_\alpha + Z_\beta\right)^2}{\left(d/\mu\right)^2}, \tag{3}$$

where the coefficient of variation, CV, is σ/μ.

For the case of the above-worked example, CV is 12.27% (= 2/16.3) and d is 1.84% of the mean (= 0.3/16.3). The calculation of the required sample number based on eq. 3 is:

$$n = \frac{12.27^2\left(1.645 + 0.8416\right)^2}{\left(1.84\right)^2} = \frac{150.5529\left(2.4866\right)^2}{3.3856} = 274.96 = 275$$

worked example for eq. (3).

4.2 Comparing lots

Note that eqs 1, 2 and 3 are one-sided tests, to be used if it is desired to detect a difference in one direction, that is, that the population mean is greater than the specification by at least d. For the detection of a difference of at least d between two populations, and assuming equal variance in the two populations, n can be calculated as:

$$n = \frac{2\sigma^2\left(Z_\alpha + Z_\beta\right)^2}{d^2}. \tag{4}$$

Consider the number of fruit that should be sampled from a consignment to determine if fruit SSC is at least 0.5% SSC higher than a second consignment. For populations of SD = 2% SSC, the required sample number can be calculated for a 5% ($\alpha = 0.05$) chance of a type I error and 80% power ($\beta = 0.2$) for type II error as:

$$n = \frac{2 \cdot 2\left(1.645 + 0.8416\right)^2}{0.5^2} = 98.9 = 100$$

worked example for eq. (4).

However, sampling almost always occurs without replacement, invalidating the sample unit independence assumption inherent in probability-based sampling designs. The impact of sampling without replacement increases

as sample size increases relative to the population size. A finite population correction (FPC) may be used to adjust estimates. As a rule of thumb, FPC is required when the sample taken is more than 5% of the population. The FPC reduces the required sample size (eq. 5) and the estimated variance (eq. 6) by a factor of $(N - n)/(N - 1)$, where N is the population size and n is the sample size:

$$n(FPC) = \frac{nN}{n + (N-1)}, \tag{5}$$

$$\sigma^2(FPC) = \frac{\sigma^2(N-n)}{(N-1)}. \tag{6}$$

For example, if in worked example 2 it was estimated that 33 fruit should be sampled. If this was for assessment of DM of the finite population of a box of 100 fruit, then eq. 5 could be used to calculate an adjusted n as:

$$n(FPC) = \frac{33\,100}{33 + (100-1)} = 25$$

worked example for eq. (5).

4.3 Estimating a population parameter

The above examples relate to the estimate of sample number to carry out a significance test. In other situations, an estimate of a population parameter such as a mean or proportion is required, within a margin of error (*e*). The required sample number for this requirement is based on the confidence interval calculation and rearranging for n. The sample size required to estimate a population mean within *e* is:

$$n = \frac{\sigma^2 \left(t_{\alpha/2,n-1}\right)^2}{e^2} = \left(\frac{\sigma t_{\alpha/2,n-1}}{e}\right)^2. \tag{7}$$

However, to obtain a *t*-statistic requires an estimate of the *df* and that requires a value for n. This 'chicken and egg' situation can be solved by using the *Z*-score as an initial solution to n and then increasing n from there until the *df* is equivalent to the solution minus 1 in eq. 7.

In the case study presented in detail in field example (i), an orchard of 494 trees had an estimated mean and SD of 88 and 82 fruit per tree, respectively. The number of trees to be counted to achieve a margin of error of ten fruit per tree with 95% confidence ($\alpha = 0.05$) can be first approximated using a *Z*-score obtained from a lookup table or from R (qnorm(0.975) = 1.961), as:

$$n = \frac{82^2\left(t_{\alpha/2,259-1}\right)^2}{10^2} = \frac{82^2\left(1.961\right)^2}{10^2} = 258.57 = 259$$

worked example for eq. (7).

Using n = 259, $t_{\alpha/2,n-1}$ is obtained from a *t*-statistic function in *R* of *qt*(0.975,259 − 1), giving a value of 1.969. The value of n can then be re-estimated as:

$$n = \frac{82^2\left(1.969\right)^2}{10^2} = 260.69 = 261$$

As 261 is greater than 259 the process needs to be repeated, this time using 261 − 1 = 260 degrees of freedom in the t-statistic calculation, resulting in n = 260.7 = 261. This is taken as the solution as the current answer for n (261) agrees with the *t*-statistic degrees of freedom (260 = 261 − 1).

Such a sample is more than 5% of the population, so an FPC adjustment is made using eq. 5 with n = 261 and N = 494, to yield a new n of 171 trees.

The sample size required for a given margin of error expressed as a percentage of the mean (PE) the calculation can be calculated as:

$$n = \frac{CV^2\left(t_{\alpha/2}\right)^2}{\left(PE\right)^2}. \tag{8}$$

Consider that if in the example above it was desired to estimate n to achieve a PE of 10% rather than an error of ten fruit/tree. In this case, using a CV of 82/88*100 = 93.2% and a *Z*-score of 1.961:

$$n = \frac{0.931^2\left(1.961\right)^2}{\left(10\right)^2} = 333.3 = 334$$

worked example for eq. (8).

This estimate can be refined using the input of a *t* score and by correcting for the size of the finite population using the FPC calculation. The outcome for the example is:

$$n\left(FPC\right) = \frac{334 \times 494}{334 + \left(494 - 1\right)} = 200$$

worked example for eq. (5).

4.4 Population proportion

When dealing with the estimation of a proportion, p, of a population with an attribute, rather than the mean, the normal approximation to the binomial

distribution is used in place of the *t*-distribution, such that $\sigma^2 \sim p(1-p)$. The required sample size from eq. 7 can be estimated as:

$$n = \frac{p(1-p)(Z_\alpha)^2}{e^2}. \quad (9)$$

Note that the closer p is to 0.5 the larger the sample size required to achieve a given margin of error. The approximation should be used only when the $n \times p$ and $n \times (1-p)$ are both greater than 5.

To calculate the minimum number of samples required to detect a difference in the proportion of fruit in the population with a specified attribute with power $(1-\beta)\times 100\%$ and significance level α, the one-sided test is:

$$n = \frac{p_1(1-p_1)(Z_\alpha + Z_\beta)^2}{d^2}. \quad (10)$$

Consider the example of sampling to detect if the rate of internal browning in apples is below the specification tolerance limit of 2% (see Field example (iii) below). To measure an attribute with an incidence level of 2% (p = 0.02) to a CI of 95% ($Z_\alpha = 1.96$) and an uncertainty of ±1% (e = 0.01) requires sampling of n fruit, following eq. 9:

$$n = \frac{0.02(1-0.02)(1.96)^2}{0.01^2} = 752.95 = 753$$

worked example for eq. (9).

A number of online calculators exist for general practitioner use. For example, the Australian Bureau of Statistics offers a resource for calculation of sample size related to the estimation of the proportion of a population belonging to a given category, for example, a survey to establish the proportion of consumers that are satisfied with a product (ABS, 2021) (https://www.abs.gov.au/websitedbs/d3310114.nsf/home/sample+size+calculator). Entering the values of 95% of confidence level, 0.02 for proportion and 0.01 for confidence interval produces a calculated value for required samples of 753.

More complicated sampling plans may be required in situations where SRS is not appropriate. For example, randomly sampling fruit across an orchard may be logistically difficult and instead a systematic or clustered sampling strategy might be preferred. However, it is important that the design used is incorporated into any calculations of the required sample size. Some online calculators on required sample number from an estimate of mean and standard deviation provide for sampling designs other than SRS (e.g. https://stattrek.com/survey-sampling/sample-size-calculator.aspx?tutorial=samp; doa 1/4/2021).

5 Estimating mean and variance

5.1 In a simple random sampling design

When each sample unit has the same probability of being selected, as in an SRS scheme, the confidence interval for the estimate of the population mean can be estimated as:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \tag{11}$$

where $\bar{x}$ is the sample mean, s the sample standard deviation, $t_{\alpha/2}$ the t-value with $n - 1$ degrees of freedom corresponding to the α level of confidence. The confidence interval describes the attribute range for which a given proportion $(1 - \alpha)$ of sample estimates of the mean will fall if sampling were repeated a large number of times. Here α is the type 1 error rate, often accepted at 0.05 (5%), corresponding to a 95% confidence interval.

Consider the example of estimation of fruit load per tree in which the number of trees in the orchard (N) is 494, s is 82 and n is 261. Using the t-statistic from the table https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm or the 'calculator' at https://stattrek.com/online-calculator/t-distribution.aspx, t is read for the 0.025 error rate, or $P(T \leq t) = 0.975$, and 260 degrees of freedom, at 1.969. Thus the 95% confidence interval for the mean is:

$$1.969 \times \frac{82}{\sqrt{261}} = 10 \text{ fruit/tree}$$

worked example for eq. (11).

As sample size, n, becomes larger relative to population size, N, the confidence interval becomes tighter. The calculation becomes:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}. \tag{12}$$

The adjusted confidence interval for the example above is:

$$10 \times \sqrt{\frac{494-261}{494-1}} = 10 \times 0.69 = 7$$

worked example for eq. (12).

These calculations assume normality; however, they are reasonably robust to departures from this assumption.

Published by Burleigh Dodds Science Publishing Limited, 2022.

5.2 In other designs

In a two-stage cluster design, a sample of PSUs is first taken and then secondary sampling units (SSU) are sampled from each PSU. The population mean is estimated as:

$$\bar{x} = \frac{K}{N}\frac{\sum_{i=1}^{k}\bar{x}_i M_i}{k} \text{ where } \bar{x}_i = \frac{\sum_{j=1}^{m_i} x_{ij}}{m_i}, \tag{13}$$

N is the total population size, K is the number of PSU's in the population, k the number of PSUs sampled, M_i the number of SSU's in the ith PSU, and m_i the number of sampled SSU from the ith PSU.

The confidence interval is then calculated as:

$$\bar{x} \pm t_{\left(\alpha/2, df=k-1\right)}\sqrt{\left(\frac{N}{K}\right)^2\left(\frac{K-k}{K}\right)\frac{s_k^2}{k} + \frac{K}{N^2 k}\sum_{i=1}^{k} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{s_i^2}{m_i}}. \tag{14}$$

where $s_k^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(M_i\bar{x}_i - \frac{\sum_{i=1}^{k}\left(M_i\bar{x}_i\right)}{k}\right)^2$, and $s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{mi}\left(x_{ij} - \bar{x}_i\right)^2$.

If all PSU's contain the same number of SSU's, M, and the same number of SSUs are sampled per PSU, m, then eq. 14 may be rewritten as:

$$\bar{x} \pm t_{\left(\alpha/2, df=k-1\right)}\sqrt{\left(\frac{K-k}{K}\right)\frac{s_a^2}{k} + \left(\frac{M-m}{M}\right)\frac{s_b^2}{Km}}, \tag{15}$$

where $s_a^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{x}_i - \frac{\sum_{i=1}^{k}\bar{x}_i}{k}\right)^2$, and $s_b^2 = \frac{1}{k\left(m-1\right)}\sum_{i=1}^{k}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_i\right)^2$

where s_a^2 and s_b^2 are the between and within-cluster variance estimates. Again, if the fraction of clusters or units within clusters sampled are small, the respective FPC's may be ignored.

Consider the example of a container of 100 pallets of 50 cartons each, with 3 cartons sampled in each of 10 pallets. Pallets are the PSU (K = 100, k = 10) and cartons are the SSU (M = 50, m = 3) (Table 1).

Using eq. 15 and s values from Table 1, the 95% confidence interval is:

$$16.6 \pm t_{\left(0.05/2, df=9\right)}\sqrt{\left(\frac{100-10}{100}\right)\frac{1.3}{10} + \left(\frac{50-3}{50}\right)\frac{1.17}{100\times 3}}$$

Table 1 Data set of a fruit attribute level for one fruit sampled from each of three cartons from each of ten pallets

Pallet	1	2	3	4	5	6	7	8	9	10	
Fruit 1	16.3	15.5	18.3	17.3	18.0	19.2	7.8	15.7	14.8	15.4	
Fruit 2	16.6	15.5	16.5	19.0	17.8	17.9	16.6	14.9	13.7	14.1	
Fruit 3	18.4	17.1	17.1	18.1	17.1	15.4	16.0	15.1	16.2	16.5	
Mean ($\bar{x}_i$)	17.1	16	17.3	18.1	17.6	17.5	16.8	15.2	14.9	15.3	$\bar{x} = 16.6$
											$s_a^2 = 1.30$
Variance (s_i^2)	1.29	0.85	0.84	0.72	0.22	3.73	0.84	0.17	1.57	1.44	$s_b^2 = 1.17$

Published by Burleigh Dodds Science Publishing Limited, 2022.

$$16.6 \pm 2.26 \times 0.35 = (15.8, 17.4)$$

worked example for eq. (15).

5.3 Multistage designs

In a multistage design employing strata with either simple random, systematic, stratified or cluster sampling, if may not be possible to select units with equal probability in each stratum. In this situation probability weights should be calculated for each unit in the sample as the inverse of the probability of selection. These weights can then be used to adjust point estimates of population parameters such as population mean and variance, with weighting based on the stratum sizes, as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{K} N_i \bar{x}_i, \tag{16}$$

where N_i and $\bar{x}_i$ are the population size and sample mean of the *i*th stratum,

$$N = \sum_{i=1}^{K} N_i$$

is the total population size, and *K* is the number of strata.

The confidence interval around this sample mean is calculated as:

$$\bar{x} \pm t_{\alpha/2, df=N-K} \sqrt{\sum_{i=1}^{K} \left(\frac{N_i}{N}\right)^2 \left(\frac{N_i - n_i}{N_i - 1}\right) \frac{s_i^2}{n_i}}, \tag{17}$$

where s_i is the sample standard deviation of the *i*th stratum and *t* is calculated using

$$\sum_{i=1}^{K} n_i - K$$

degrees of freedom. Note that

$$\left(\frac{N_i - n_i}{N_i - 1}\right)$$

is the FPC (eq. 5).

Consider the data from Table 1 with the modification that there are only ten pallets in the container to provide an example of a stratified design. Samples are taken from all pallets in the population, so a given pallet is treated as a strata and carton becomes the PSU. The calculation of the 95% confidence interval becomes:

$$16.6 \pm t_{\left(0.05/2, df=30-10\right)} \sqrt{\sum_{i=1}^{k=10} \left(\frac{50}{500}\right)^2 \left(\frac{50-3}{50}\right) \frac{s_i^2}{3}}$$

$$16.6 \pm 2.09 \sqrt{\frac{1}{100} \times \frac{47}{49} \times \frac{11.7}{3}} = \left(16.2, 17.0\right)$$

worked example for eq. (17).

If restricted by the number of samples possible it is often better to sample as many PSUs as possible at the expense of the lower number of units per PSU, assuming most variation occurs at this level. Conversely, if it is of interest to quantify the within – and between – cluster variance, sampling of more units per cluster is required.

In most cases, the calculation of variance estimates is not trivial in multistage designs. Popular methods for variance estimation include the linearisation methods, resampling techniques such as jackknife, balanced repeated replication and bootstrap. The simplest of the replication methods is the jackknife. In this approach, *k* jackknife replicates are created, where *k* is the number of PSUs sampled. Each replicate removes a unique PSU from the sample, adjusts the weights accordingly, and calculates the statistic of interest. The variance is then estimated by how different the replicate statistics are from the full sample statistic. To calculate the variance of the mean, the weighted mean is first calculated as:

$$\overline{x} = \frac{\sum_{j=1}^{k} \sum_{l=1}^{m} w_{jl} x_{jl}}{\sum_{j=1}^{k} \sum_{l=1}^{m} w_{jl}}, \tag{18}$$

where *k* is the number of PSUs and m is the total number of units in the PSU. For non-stratified samples, the jackknife coefficient used to adjust the replicate weights is defined as:

$$r = \frac{k-1}{k}. \tag{19}$$

See SAS (2021) for the calculation used when stratification is used. The *i*th jackknife replicate mean is then calculated as:

$$\hat{x}_i = \frac{\sum_{j=1, j\neq i}^{k} \sum_{l=1}^{m} w_{jl} x_{jl}}{\sum_{j=1, j\neq i}^{k} \sum_{l=1}^{m} w_{jl}}, \tag{20}$$

and the jackknife standard error of $\overline{x}$ is then:

$$s = \sqrt{\sum_{i=1}^{k} r\left(\hat{x}_i - \hat{x}\right)^2}, \tag{21}$$

$$\text{where } \hat{x} = \frac{\sum_{i=1}^{k} \hat{x}_i}{k}. \tag{22}$$

Owing to the complexity of variance estimates it is recommended to use specialist software such as STATA, SAS, or R (survey package) to analyse data collected under such schemes.

The next section of the chapter considers some specific case studies to position sampling schemes in given situations.

6 Case studies

6.1 Pre-harvest estimation of crop load

6.1.1 The challenge

Estimation of crop load in the field, for example, the number of fruit on the trees in an orchard, is typically achieved by counting of fruit from a sample of trees. A set number of trees are typically counted in commercial practice, with

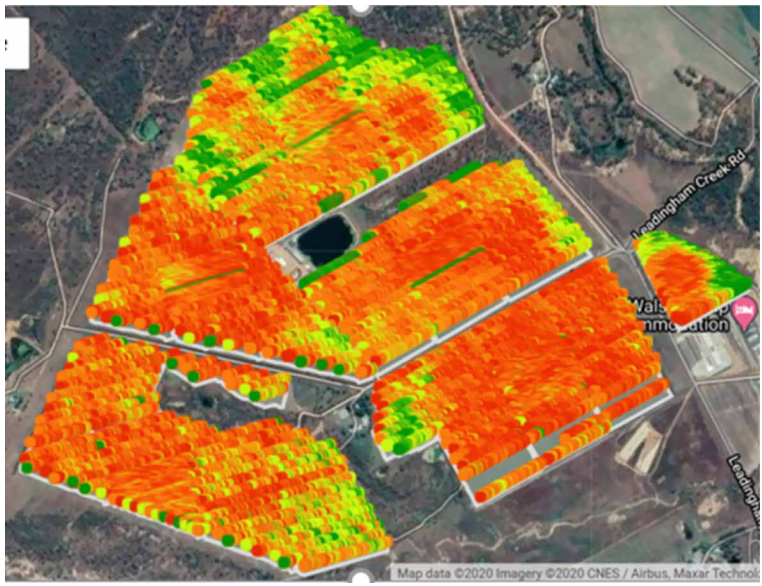


Figure 3 Heat map (green/yellow/orange/red, low to high) of fruit load (#fruit/tree) across an orchard from machine vision assessment.

trees selected on a line through the orchard for operator convenience. This methodology is flawed, in terms of both sample number and location. If tree-to-tree variation is high, more sample units (trees) should be assessed to reduce the uncertainty of the estimate. If tree yield is spatially heterogeneous across the orchard (Fig. 3), sampling patterns should be informed by knowledge of that spatial variation. An easily assessed attribute that is correlated to fruit load is required to stratify orchards. Remotely assessed indices like canopy normalised difference vegetation index (NDVI) have been used, for example, by Wulfsohn et al. (2019), Rahman et al. (2018) and Anderson et al. (2019), although the correlation of vegetation index to fruit load is sometimes weak.

6.1.2 Required sample number in SRS

In Table 2, data is presented of the fruit load on six trees chosen randomly within each of three equal-sized canopy NDVI classes (a stratified design) in each orchard. The sample number required for an estimate of the mean using SRS at the stipulated confidence level, and %error was calculated using the mean and SD of this preliminary sample by application of eq. 8, with adjustment for the finite sizes of the populations using eq. 5.

In farm practice, the undertaking of a preliminary sampling to estimate population SD to achieve a desired error (e.g. <10% of mean) can be replaced by an on-the-go estimate. For example, Wulfsohn et al. (2019) describe a tablet-based app for entry of data, allocation of random starting points in a stratified sampling design and calculation of required sample number (Pronofruit 2021).

Table 2 Data on tree number and mean, SD and CV on fruit number per tree for 18 trees in each of 10 mango orchards, with calculated sample *n* and *n** required for a 95% confidence and an error of 10% in estimation of average fruit number per tree, where *n** is after adjustment for population size

Orchard	*N* (#trees)	Mean (fruit#/tree)	SD (fruit#/tree)	CV (SD/M)	*n* (#trees)	*n** (#trees)
1	469	88	82	93	334	195
2	486	259	102	39	60	53
3	1017	240	160	67	171	146
4	1100	80	34	43	69	65
5	224	59	36	61	143	87
6	1205	97	65	67	173	151
7	1091	201	55	27	29	28
8	1818	106	51	48	89	85
9	1176	77	61	79	241	200
10	1117	85	40	47	85	79

Data from Anderson et al. (2019).

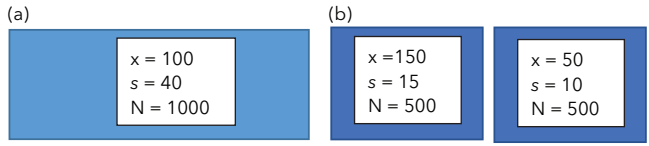


Figure 4 Hypothetical case of (i) estimate error achieved for a count of a set number of trees, and (ii) number of trees to be counted to estimate average tree fruit load of an orchard, *n*, for a set estimate error, *e*, and probability, *P*. Estimates are calculated using average number of fruit/tree, *x*, the standard deviation of the number of fruit/tree, *s*, the number of trees in the orchard, *N*, through application of eqs 5, 7 and 13. The required finite populate adjusted n is denoted *n**, and adjusted *e*, *e**. Two conditions are considered: (a) SRS of the entire orchard, and (b) a stratified design with SRC within two classes.

6.1.3 Other sampling designs

The required sample number can be decreased if sub-populations with a lower variation that the total population can be identified through stratified, systematic or cluster sampling, as appropriate. For example, consider a hypothetical orchard with areas of high- and low-yielding trees separated spatially. SRS sampling of 50 trees over the entire orchard results in an estimate error of 11 fruit per tree, while sampling of 50 trees in two sub-populations results in a decrease in measurement error (Fig. 4).

o achieve a measurement error of 5 fruit/tree at 95% probability, the sample number could be reduced from 246 to (33 + 16 =) 49 trees.

Alternatively, using the right-hand side of eq. 17 for a stratified design to estimate the population mean from a sample of 50 trees (25 per strata) gives e = 3.5:

$$e = t_{0.025,df=50-2}\sqrt{\left(\frac{500}{1000}\right)^2\left(\frac{500-25}{500-1}\right)\frac{15^2}{25}+\left(\frac{500}{1000}\right)^2\left(\frac{500-25}{500-1}\right)\frac{10^2}{25}} = 3.54$$

Thus, the sample number can be reduced from 246 to 28 trees for a measurement error of 5 fruit/tree:

$$e = t_{0.025,df=28-2}\sqrt{\left(\frac{500}{1000}\right)^2\left(\frac{500-14}{500-1}\right)\frac{15^2}{14}+\left(\frac{500}{1000}\right)^2\left(\frac{500-14}{500-1}\right)\frac{10^2}{14}} = 4.89$$

This decreased sampling effort illustrates the benefit of a more efficient sampling strategy compared to SRS.

A systematic-uniform-random sampling strategy is advocated by Wulfsohn et al. (2019) to reduce counting effort over traditional 'random' sampling and to provide an advantage in field convenience, in terms of locating trees. This procedure involves manual fruit counts of systematically selected row, tree, branch and branch segments, with the sampling unit defined by the crop and canopy architecture. The periodicity of sampling requires prior knowledge of the source of variation (within vs between trees). The method yields a high number of small units distributed uniformly in the three dimensions of tree canopies. The low counts per sample unit are considered important to reduce human errors in counting.

6.2 Optimising time of harvest

6.2.1 New Zealand kiwifruit example

Kiwifruit growers in New Zealand are required to ensure industry-set 'maturity clearance' standards are met before harvesting a crop. Rather than the use of a set attribute level that the average of the sample must exceed, the maturity metric is a set percentile point. The maturity clearance standard for the gold-fleshed Zespri™ SunGold Kiwifruitis is the 90th percentile for colour (90% of the sample must be below a specified hue value) and the 30th percentile for DM (70% of the sample must have a DM value greater than a specified threshold). Additionally, strict and detailed sampling protocols must be followed in the selection of fruit used for the maturity clearance test, in terms of sample number and sampling strategy.

'Maturity areas' (MAs) of an orchard are selected on the basis of providing homogenous growing conditions, consistent with respect to terrain and climate factors, which will minimise variance within that scale. The area is typically around 1 ha and cannot exceed 4 ha, with a typical yield > 12 000 tray equivalents or 360 000 average-sized fruit. Perhaps surprisingly then, the typical maturity clearance sample from an MA is only 90 fruit, only 0.025% of the typical fruit number.

The required sample number to achieve a desired measurement uncertainty and probability is a function of population SD, not population size (cf. eq. 1). In the New Zealand kiwifruit industry, it has been empirically established that there is little to no benefit to have sample sizes larger than 90 (Fig. 5), at least in comparison to the increased complexity and costs that would be incurred with larger sample sizes.

How should the required sample of 90 fruit be acquired from a 1 ha field containing 360 000 fruit? The development of a sampling strategy to achieve a true representation of the heterogeneity is a key challenge with maturity testing given fruit-to-fruit variation is large on orchards at all scales, be it within single

Published by Burleigh Dodds Science Publishing Limited, 2022.

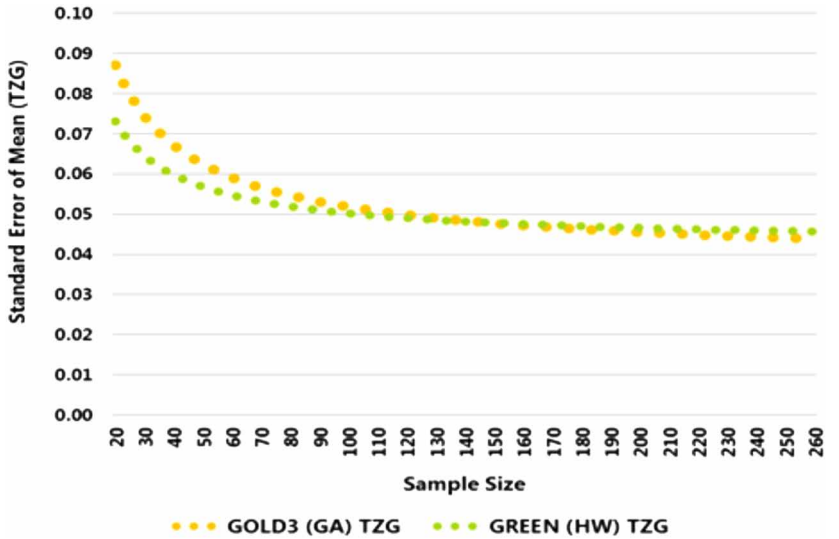


Figure 5 Standard error of mean DM (% FW) versus sample size, with regard to the Taste Zespri Grade (TZG) standard (Zespri, 2018). Gold dots represent Zespri™ SunGold Kiwifruit; green dots represent Zespri™ Green Kiwifruit.

fruit, between fruit in a single plant, between plants, within orchard blocks and between blocks. The kiwifruit sampling protocol involves a stratified multistage strategy of first mapping the entire MA into a spatially representative grid sampling pattern of 90 sampling vines. The area of each sampling vine is then divided into three representative lanes (i.e. fruit canes close to the leader line, halfway out along the canes and then close to the end of the canes) and sampled on a rotational basis through the sequence of sampling vines. Low-hanging fruit, which are easily accessed and would tend to be over-represented in unsupervised human sampling and sun-exposed upper canopy fruit, tend to be outliers in attribute levels. Fruit are therefore taken from within a ±10 cm zone of the middle canopy, enforcing consistency between the designated strata pattern. The importance of proper sampling is underlined by the fact that only fully independent and audited operators are authorised to do the sampling.

These kiwifruit sampling protocols are continually being updated and refined, on a near yearly basis, as more is learnt about orchard and vine variability. Change in crop consistency due to changes in genetics, agronomic practice or growing conditions requires a change in sampling effort. For example, the 'improved' Gold cultivar requires greater sampling for the same certainty of measurement (Fig. 5), and inconsistent pollination can increase crop variability, requiring greater sampling effort for the same certainty of estimation.

6.3 Post-harvest: in-line sorting statistics and the impact of measurement error

6.3.1 Measurement error in packlines

The fresh produce packline represents a point in the supply chain where all units can be individually assessed using a range of sensors. Weight assessment is achieved with relatively high accuracy, approximately ±1 g, with measurement of fruit on conveyors moving at around 1 m s^{-1}. Machine vision allows discrimination of fruit on the basis of external features such as colour, shape and surface defects like scars, blemishes and bruises. Near-infrared spectroscopic (NIR) sorting systems allow sorting on internal fruit quality factors such as sweet taste (SSC, DMC) and/or the presence of defects (internal browning, cavities). The machine vision and NIR sorting systems are generally used to provide binary sorting classifications of fruit, into acceptable and unacceptable grade categories.

However, the training of a machine vision method to recognise specific external defects, and the training of models based on NIR spectra for assessment of internal attributes of fruit assessment, is a time-consuming and difficult exercise. The predictive models can be prone to relatively large errors. The consequence is that these newer sorting operations should be operated with a system to deliver regular and current information on the error or misclassification rates in sorting. The error rates are essential knowledge for maintaining the integrity of the sorter output, whether it is in maintaining high-grade category standards and/or minimising losses to the low-grade category. Such a quality control system requires a sampling strategy in the choice of samples.

For example, the measurement error associated with NIR spectroscopy is often reported as the root mean square of error of prediction (RMSEP),

$$RMSEP = \sqrt{\frac{\sum_1^n (p-a)^2}{n}}, \tag{23}$$

where p is the predicted value, a is the actual or reference method value and n is the number of samples. The RMSEP encompasses prediction accuracy, as expressed by bias, the average difference between actual and predicted values as well as prediction precision, as:

$$RMSEP = \sqrt{bias^2 + SEP^2}, \tag{24}$$

where SEP is bias-corrected RMSEP, i.e.

$$SEP = \sqrt{\frac{\sum_1^n (p-a-bias)^2}{n}}. \tag{25}$$

In practice, SEP is reasonably stable for established (multi-year) NIR spectroscopic-based models of attributes that are commonly assessed in commercial practice (e.g. >0.5% FW for DM content and 0.5% w/v for SSC). However, unacceptable biases (>1% FW on DM) on prediction can still occur with new incoming populations. This bias can be associated with an instrument or population changes. Changes in temperature of fruit or instrument can be accommodated in the modelling process, there being a predictable impact on water absorption peaks and detector sensitivity, respectively. Other population changes are less predictable. For example, growing conditions may affect fruit cell size, skin thickness and composition, with impact on a NIR-based prediction, primarily as bias.

Bias can be quantified by measuring the average difference between predicted and actual values on a small sample set. Precision on an estimate of bias, s_{bias} can be calculated as:

$$s_{bias} = \frac{\sqrt{\sum_1^n (p-a)^2}}{n} = \frac{RMSEP}{\sqrt{n}}. \quad (26)$$

Note that as sample size, n, appears as a divisor in eq. 26, it does not have to be very large to achieve good bias precision.

Bias can be 'simply' accommodated in further predictions. This correction is central to commercial viability for NIRS applications, avoiding the need for expensive re-development of multivariate models. However, there is an art involved in judging how often to check for the need for adjustment.

6.3.2 Packline sampling

The standard statistical approach to understanding classification errors involves the assessment of representative fruit samples from both the acceptable and the unacceptable output bins of the sorting operation. As with all sampling, the need for representative sampling from the output bins is paramount, requiring care to avoid the introduction of sampling bias. For instance, randomly sampling from only the top layers of an output bin will result in the selection of fruit representative of those recently processed by the sorter, and not of the whole consignment. Fruits presented to the grader generally have a heterogeneity correlated with in-orchard trends that match fruit harvesting order. For example, if a number of trees producing high-quality fruit are harvested together, it is likely that these fruits will also present together to the sorter. There will be fewer unacceptable fruit graded as acceptable than in the overall population. A systematic sampling approach could be employed, with the grader directing every nth sample to

an assessment bin, rather than a collection of fruit from the top layers of fruit in the output bin at a single point in time. Alternatively, a stratified design could be employed, given knowledge of orchard variation in the attribute of interest.

6.3.3 Sorting statistics: apple browning and kiwifruit DM examples

Sorting involves a pass/fail binary classification, termed positive (P) and negative (N) classes. The terms P and N are typically used in the sense that a sorter is used to find and remove unacceptable fruit as 'positive' detections and otherwise to report 'negatively' or ignore the acceptable fruit. Fruit in the P grade will include both true positives (TP) and false positives (FP), the latter being acceptable fruit incorrectly assessed as unacceptable. Fruit in the N grade will consist of true negatives (TN) and false negatives (FN), the latter being unacceptable fruit falsely classified as acceptable. Examination of representative fruit samples from the respective output bins enables calculation of the number of FP, also known as a false alarm or a type I error and FN, also known as a miss or a type II errors. A common practice is to lay out the statistical results of the samplings as a confusion matrix, also known as an error matrix (Fig. 6).

Sometimes the classification is based on the measurement of a discrete or continuous variable, rather than a direct binary classification. Consider the case of a sorting technology, which non-invasively assesses the severity of internal browning in an apple on a 1 to 5 scale, in which scores of 1 and 2 are considered acceptable and 3–5 as unacceptable (Khatiwadi et al., 2016). The incidence of FP and FN results will change with the proportion of P and N in the incoming population, with the rate of measurement error and with threshold value or cut-point around which the sorting decision is made (Fig. 7). For example, as the mean score of a population decreases, the incidence of defect fruit accepted (FN/P) will increase for a given sorting threshold. Decreasing the acceptable threshold will decrease FN/P but will increase the incidence rate of acceptable fruit rejected (FN/P). Optimisation of the sorting operation requires knowledge of the instrument measurement error and the mean and SD of the defect in the incoming population.

n = 227	**Predicted acceptable**	**Predicted defective**
Actual acceptable	50	21
Actual defective	0	156

Figure 6 Example of a confusion matrix: results of a non-invasive sorting operation (using a threshold of 2, see Fig. 7) on 227 fruit to separate fruit with internal browning defect from acceptable fruit, compared to actual, as assessed from visual inspection of cut fruit.

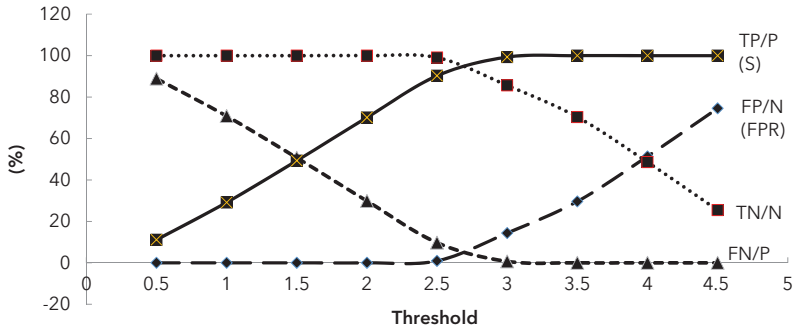


Figure 7 Example of impact of change in threshold level in a sorting operation on apple internal browning in terms of % of good fruit correctly accepted (*TN/N*) and rejected (*FP/N* = false-positive rate), and of defect fruit accepted (*FN/P*) and rejected (*TP/P* = sensitivity), for a population of mean score 3.1 and SD 1.4 (data of Khatiwadi et al., 2016).

However, sorting statistics and decisions are often made on the basis of only very small subsamples from the output bins, without accurate knowledge of the incoming population in terms of the number of actual good (*N*) and poor (*P*) fruit. What is generally known with good accuracy is the recovery rate of the sorting operation, the volume fraction of outgoing fruit to incoming fruit ($N/(P + N)$). With that parameter and estimates of the false positive and negative proportion in the sorted bins, the sensitivity and FPR can be calculated as:

$$\text{Sensitivity} = \frac{1}{\left(1+\frac{PnN}{1-NnP}\right)\left(\frac{R}{1-R}\right)}, \tag{27}$$

$$\text{False Positive Rate} = \frac{1}{\left(1+\frac{NnP}{1-PnN}\right)\left(\frac{1-R}{R}\right)}, \tag{28}$$

where *PnN* and *NnP* are sample estimates of the fractional number of fruit incorrectly sorted into the good (*N*) and poor (*P*) classes, respectively, and *R* is the total recovery rate in terms of the volume of good fruit.

As in much of statistics, there is seldom one single best summary statistic or graph to represent all aspects of a sorting operation. Common examples include accuracy (ACC), F-score (F1), area under curve (AUC) and the Matthews correlation coefficient (MCC). A receiver operating characteristic (ROC) curve is commonly used to understand sorting efficiency, if it can be conveniently assembled. It is a plot of the true positive rate (*TP/P*; also known as the sensitivity) against the false-positive rate (*FP/N*; equivalent to 1 − specificity) (Ooms et al., 2010). It can be time-consuming to generate as it requires varying the threshold

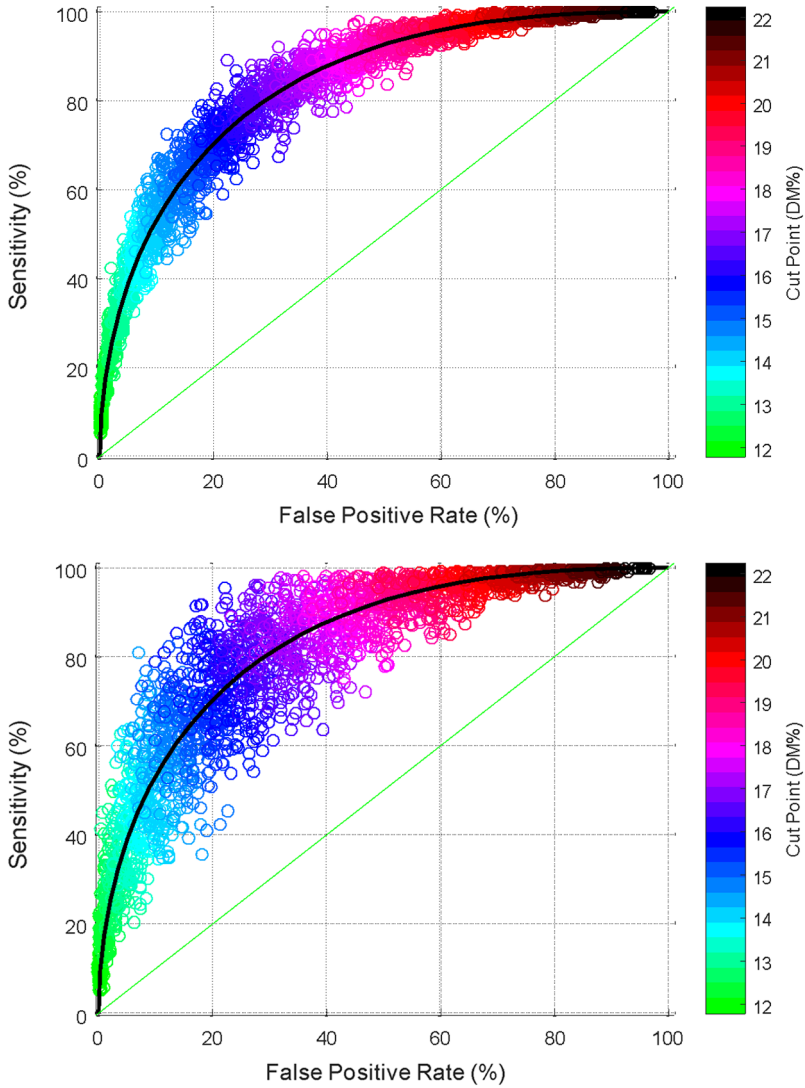


Figure 8 ROC (Sensitivity against false positive rate) plots for a simulation of NIR spectroscopy sorting of kiwifruit to a DM specification 16.1% FW. The statistics were generated by random subsampling of 50 (top panel) or 200 (bottom panel) fruit from both the high- and low-DM sorted output bins. The colour scale represents the variation with change in the selected sorting threshold value (colour bar).

or cut-point value across a large fraction of the incoming population range. Some simulated sorting data are shown in the ROC format in Fig. 8, where the scattered points represent various randomised re-samplings of the output bins, at various threshold levels, and with the black curves being a calculated

running averaged ROC curve in each case. The curves illustrate the trade-off between the success of the sorting operation in removing actual defective or poor fruit (*TP*/*P*; sensitivity) and the cost that must be simultaneously met in removing good fruit (*FP*/*N*; false-positive rate FPR). The further the curves in a ROC plot are towards the upper left-hand corner, of high sensitivity and low FPR, the more accurate the sorting operation.

There are alternative framings of the classification problem, such as that described by signal detection theory (SDT; Bollen and Prussia, 2009). SDT theory uses similar concepts to the classification statistics described above but then narrowly frames the problem as the detection of one distinct 'signal' population within another, for instance, of two distinctly different normally distributed populations of differing means and standard deviations. The reframing enables the specification of new metrics from the output sample data, such as the detectability *d'* (the difference between the means) and the setpoint criterion *S* (an optimising cut-point), both of which can be used to conveniently compare the performance of different sorting operations. However, the SDT approach is valid only for those fruit sorting applications where there are distinctly different sub-populations to segregate, such as in detecting a small diseased fruit population within a larger healthy population.

In some applications, such as with NIR-based sorting on internal fruit quality, the goal is simply to separate a low or high fraction from a single population distribution. In that case, there are not two separate populations to distinguish, rather just one that has to be split, and the theoretical modelling required in that circumstance involves truncated distributions, split around the chosen cut-point. As analytical methods of analysis do not exist for those circumstances, the sorting scenario needs to be modelled numerically, for example, with Monte Carlo methods (e.g. Harrison, 2010).

Given knowledge of the measurement error (i.e. SEP) for a NIR spectroscopy-based prediction of an attribute and knowledge of attribute distribution (mean and SD) in a population, it is possible to model ROC curves, as demonstrated in Fig. 8 for a simulated DM segregation of kiwifruit. The simulation exercise involved the calculation of a recovery rate and resampling of 50 fruit from each of the good and poor bins, to calculate the *PnN* and *NnP* parameters for a sorting threshold of 16.1% FW. The data were generated from a simulated fruit DM population (of the normal distribution with mean 17% FW and variance 2% FW), with added NIR prediction noise (normal with mean 0% FW and variance 1% FW). The exercise revealed considerable spread in the reported data, which must be due to sampling variation (Fig. 8). Further simulation studies showed clear benefits of increasing sample sizes for the good and poor bins (Fig. 8, bottom panel). In this example, the average spread around the true or mean ROC curve in the simulations decreased from near 10% to below 5% as the sample size increased from 50 fruit to 200 fruit.

6.4 Post-harvest: sampling for biosecurity assessment

6.4.1 Sample selection

Sampling inspections of harvested and packed fruit are often mandated by industry regulation to ensure fruit lots meet set quality standards. The sampling practices required to provide assurance that a consignment meets pest and disease standards are relatively well developed relative to those for eating quality assessment, because they are critical to international trade, and the biosecurity risks involved in the failure to detect are huge.

The International Plant Protection Convention provides sampling standard ISPM 31 (IPPC, 2020). The standard involves taking a subsample of a consignment for visual inspection, with the decision of acceptability dictated by the number of positive samples. The size of the subsample is chosen based on the desired degree of confidence that the pest or disease will be picked up for a given rate of infestation. ISPM 31 provides an example case where 'at a 95% confidence level, not more than 0.5% of the units in the consignment are infested'(MAF Biosecurity New Zealand, 2008), where a 'unit' is typically an individual fruit. The sample size required to achieve this confidence is often modelled as a binomial distribution and assuming a given acceptance level, that is, infested units allowed in the subsample.

Let X be the number of infested units in the subsample, then X will be binomally distributed ($X \sim Binom(n,p)$), where n is the subsample size and p is the proportion of infested units in the whole lot.

The sample size required can then be found by setting p = detection level (0.005 in the above statement), and finding the minimum n such that ($X > 0$) = 95% for a 95% confidence level. This is equivalent to solving for n such that $Pr\,(X = 0) = 5\%$, which results in a value of 598 units (see calculation below). Sampling of 600 units is mandated in many phytosanitary inspection protocols with the lot failing inspection if one or more sampled units are out of specification. This equates to a 95% confidence in rejecting an infestation level of 0.5%, that is, a proportion of 0.005.

Equation 9 can be used to get an approximate solution by setting $e = p = 0.005$ and $\alpha = 0.05$:

$$\frac{p(1-p)(Z_\alpha)^2}{e^2} = \frac{0.005(1-0.005)(1.645)^2}{0.005^2} = 539$$

An exact solution can be found by setting the probability of no infested units in the sample to α, that is:

$$\alpha = (1-p)^n$$

Published by Burleigh Dodds Science Publishing Limited, 2022.

which can be rearranged to solve for n:

$$n = \frac{log(\alpha)}{log(1-p)} \tag{29}$$

$$n = \frac{log(0.005)}{log(1-0.005)} = 598$$

worked example for eq. (29).

This can also be computed using the cumulative density function of the negative binomial distribution. This can be calculated in *R* using the qnbinom function. In the above example qnbinom(0.95,1,0.005) = 597, which is the upper 95% limit on the number of clean samples before one infested sample is observed. That is, the total samples needed are 597 + 1 = 598. If the population size was 1000 with 5 infested (infestation level of 0.005) and 995 clean units, the required sample size can be calculated using the negative hypergeometric

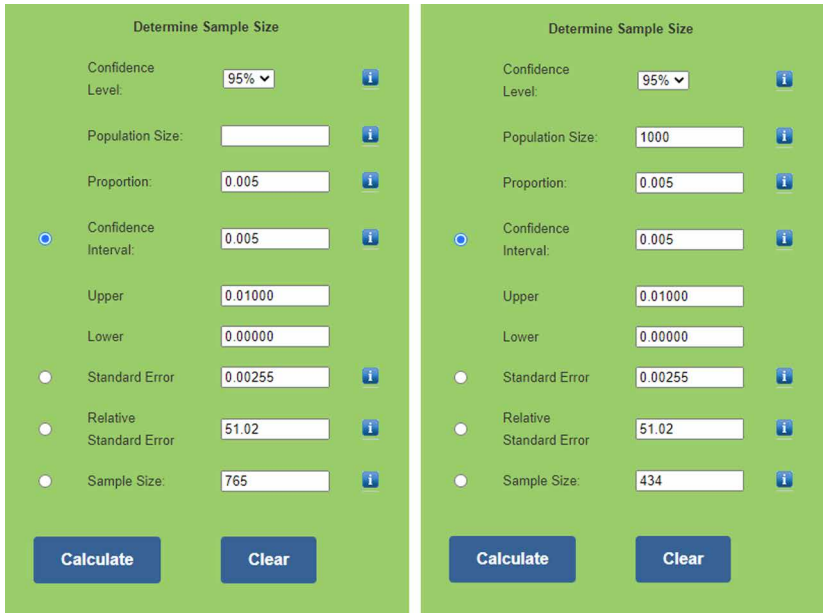


Figure 9 ABS sample size calculator (https://www.abs.gov.au/websitedbs/d3310114.nsf/home/sample+size+calculator) for the case of detection of a defect at 0.5% incidence rate at a 95% confidence interval, for an infinitely large population (left panel) and a population of 1000 units (right panel).

distribution. The *R* function qnhyper(0.95,995,5,1) from the extraDistr package gives 450 samples. Note that the qnhyper function gives the total sample size as the solution so does not require adding the minimum number of infested units in the sample before the sample fails inspection (1 in this example).

The Australian Bureau of Statistics calculator (ABS, 2021), which is based on eq. 9, can also be used with this example (Fig. 9), with the limitation that the calculator only allows for 95% or 99% confidence interval. As the test in the example is one-sided (infestation < 0.005), the previous estimates were based on a 2.5% chance (not 5%) that the sample will be clean if the population has 0.5% infestation.

This calculation assumes that the visual inspection will always detect any infested units (fruit) included in the subsample. Relaxing this assumption requires adjustment of the *P* parameter. For example, if visual inspection correctly classifies an infested unit 90% of the time, then in the above example $P = 0.005 \times 0.9 = 0.0045$. Using eq. 29, this results in a requirement for a sample of

$$\frac{\log(0.05)}{\log(1-0.0045)} = 665$$

units to achieve a 95% confidence that a lot with an infestation rate of 0.5% will fail inspection.

ISPM 31 tabulates the required sample size in the context of sample detection level and confidence level (Table 3).

A requirement is placed on the above estimates that the population size is large relative to the sample size. When the sample size is small (i.e. <5% of

Table 3 Sample size required for detection of a given infestation rate in a large lot (as defined by the lot size being at least 20 times larger than the sample size) at a given confidence level (from ISPM 31)

Infestation rate (%)	Confidence level (%)	Sample size (units)
5	90	45
5	95	59
5	99	90
5	99.9	135
0.1	95	2995
0.5	95	598
1	95	299
2	95	149
5	95	59
10	95	29

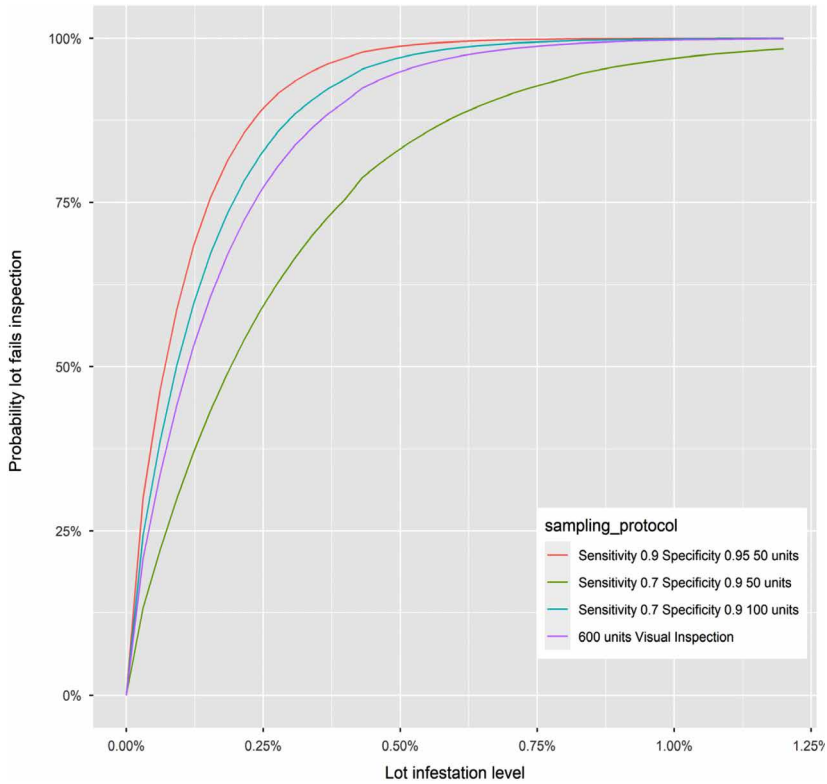


Figure 10 Probability a lot failing inspection under various sampling protocols. The standard 600-fruit visual inspection curve without accounting for the finite population are given as a reference. Calculations were based on a lot of 10 000 units with varying infestation levels. The simulation assumed grading of 1000 units combined with visual inspection of a 50 or 100 unit subsample of lots graded as positive.

N), the hypergeometric distribution should be used in place of the binomial distribution. Defining the number of infested units included in the visual inspection sample as *X*, then $X \sim (N,K,n)$, where *N* and *n* are the population and sample sizes, respectively, and *K* is the total number of infested units in the lot. A 95% confidence of detection for a lot size of 1000 with 5 infested units (0.5% incidence) is achieved with a sample of 450 units (see also Fig. 10).

6.4.2 Adding sorting technology to biosecurity assessments

Currently, phytosanitary inspections are undertaken manually. It is interesting to consider the impact of the use of a non-invasive assessment technology installed on a fruit grading platform. Such a technology could be used to assess all the units of a population or a large subsample of the population.

This advantage is counterbalanced by the disadvantage that detection of the infested fruit by the non-invasive method would probably be less reliable than a destructive method involving human assessment.

Two important measures of this performance are sensitivity and specificity. Sensitivity is the probability that the grading will correctly detect an infested fruit. Specificity is the probability that the grader will correctly classify a clean unit. It is desirable to have both as high as possible, but this can vary by technology, the algorithm used for prediction and also the severity of infestation. For example, sorting of a consignment of fruit with mild disease symptoms will suffer a lower sensitivity if a larger proportion of infested units are incorrectly classified as clean.

The assurance provided by automated grading compared with manual visual inspection is influenced by these two measures. Lower sensitivity is often less of an issue as the increased sample size for the automated grading will offset this drawback, and it will outperform a smaller sample assessed offline or online by manual visual inspection. Specificity, however, could pose a significant issue if not 100%. Specificity less than 100% means that the inspection is at risk of false positives, with this rate amplified by the high number of units graded in terms of the number of false positives encountered. A specificity of 100% is generally unrealistic and so caution should be used to reject lots based solely on a small number of units classified by the grader as infested. An additional subsample of the units rejected by the grader should then be taken for visual inspection. This subsample would be smaller than the 600-fruit sample required if the decision was based solely on visual inspection of randomly sampled units as the infestation rate should be higher in the rejected units than in the population.

Ignoring the finite population, the standard calculations of sample size can be based on the expected infestation rate of the rejected unit from the grader rather than the whole-lot infestation rate if the full lot is graded. For example, a grader with 90% specificity and 80% sensitivity grading units used in the assessment of a population with a 0.5% incidence rate would expect to achieve a precision (correctly detected positives/all detected positives) in the identification of infested units of:

$$\frac{\text{true positives}}{\text{true positives} + \text{false positives}} = \frac{0.8 \times 0.005}{0.8 \times 0.005 + (1-0.9) \times (1-0.005)}$$

$$= 0.039$$

From eq. 29, this would require a sample of only

$$\frac{\log(0.05)}{\log(1-0.039)} = 76\,,$$

Published by Burleigh Dodds Science Publishing Limited, 2022.

rather than 600, units to achieve 95% confidence that a lot with an infestation rate of 0.5% will fail inspection. Simulations of full lot grading, at varying sensitivity and specificity rates, suggest a good advantage in reduced sampling sizes for at-line tests (Fig. 10).

7 Conclusion

New measurement technologies are facilitating new approaches for the improvement of safety and quality in agri-food supply chains. However, measurement uncertainty and choice of sampling strategy can influence the effectiveness of sampling outcomes. This chapter provides a sampler of calculations of population statistics, required sample sizes and approaches to sampling strategy and provides an insight into the complex considerations that need to be undertaken to ensure that the results from the sampling exercise are representative and without bias. It is essential to consider the degree of heterogeneity of the product population itself and whether a given defect is equally distributed through the product population or is a discrete issue (specific grower in a consignment from multiple growers or an area of an orchard in a whole crop harvest) that has its own pattern of distribution within the sub-population of a whole consignment. While examples from the fresh produce sector are given, the themes explored in this chapter are of relevance to the student and practitioner operating in agri-supply chains more generally.

8 Where to look for further information

The topics covered in this chapter are of practical relevance to fruit value chains and thus relevant techniques are embedded into the operation of organisation operating through the value chain. By way of examples, reference has been made in this chapter to commercial operations such as Geco Enterprises (San Vicente TT, Chile), a company involved in crop load estimation, and Zespri (New Zealand), a company involved in kiwifruit marketing. Organizations involved in quality control inspections, particularly biosecurity inspection, are also practitioners of the art of sampling. As recommended in section 2, documentation on sampling for fruit specifications as provided by the United States Department of Agriculture (USDA, 2021) is well presented. For statistics theory, the text by Thompson (2012) is recommended.

9 Acknowledgements

We thank Dvoralai Wulfsohn of Geco Enterprises Ltda, San Vicente TT, Chile, for critical comments on the manuscript.

10 References

ABS (2021). Sample size calculator. Available at: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/sample+size+calculator (accessed 30/1/2021).

AMS (2020a). Banana inspection instructions. Available at: https://www.ams.usda.gov/sites/default/files/media/Bananas_Inspection_Instructions%5B1%5D.pdf (accessed 2/12/2020).

AMS (2020b). Lemon inspection instructions. Available at: https://www.ams.usda.gov/sites/default/files/media/Lemons_Inspection_Instructions%5B1%5D.pdf (accessed 2/12/2020).

Anderson, N. T., Underwood, J. P., Rahman, M. M., Robson, A. and Walsh, K. B. (2019). Estimation of fruit load in mango orchards – tree sampling considerations and use of machine vision and satellite imagery. *Precision Agriculture* 20(4), 823–839. https://doi.org/10.1007/s11119-018-9614-1.

Bollen, A. F. and Prussia, S. E. (2009). Sorting for defects and visual quality attributes. In: Florkowski, W. J., Shewfelt, R. L., Brueckner, B. and Prussia, S. E. (Eds), *Postharvest Handling* (2nd edn.). Academic Press: Boston, pp. 399–420. https://doi.org/10.1016/B978-0-12-374112-7.X0001-7.

Citrus Australia (2003). Measuring crop load. Available at: https://citrusaustralia.com.au/wp-content/uploads/Fruit-Size-Management-Guide-Part-2.pdf (accessed 7/2/2021).

DAWR (2021). Guideline: inspection of horticulture for export. Available at: https://www.agriculture.gov.au/sites/default/files/sitecollectiondocuments/biosecurity/export/plants-plant-products/plant-exports-manual/guideline-inspection-horticulture-export.docx (accessed 7/2/2021).

Farmsoft (2020). Quality control fresh produce. Available at: https://www.farmsoft.com/traceability/quality-control-fresh-produce (accessed 20/11/2020).

Follett, P. A. and Hennessey, M. K. (2007). Confidence limits and sample size for determining nonhost status of fruits and vegetables to tephritid fruit flies as a quarantine measure. *Journal of Economic Entomology* 100(2), 251–257. https://doi.org/10.1603/0022-0493(2007)100[251:classf]2.0.co;2. PMID: 17461044.

Freshmark (2020). Produce specification, apple. Available at: https://freshmark.com.au/wp-content/uploads/2015/05/Apple.pdf (accessed 7/3/2021).

Harrison, R. L. (2010). Introduction to Monte Carlo simulation. *AIP Conference Proceedings* 1204, 17–21. https://doi.org/10.1063/1.3295638. PMID: 20733932; PMCID: PMC2924739.

International Plant Protection Convention (2020). Standard ISPM 31. Available at: https://www.ippc.int/en/publications/83473/ (accessed 7/1/2020).

Khatiwadi, B. P., Subedi, P. P., Hayes, C., Cunha Carlos, L. C. C. and Walsh, K. B. (2016). Assessment of internal flesh browning in intact apple using visible-short wave near infrared spectroscopy. *Postharvest Biology and Technology* 120, 103–111.

MAF Biosecurity New Zealand (2008). Importation and clearance of fresh fruit and vegetables into New Zealand. Available at: https://www.mpi.govt.nz/dmsdocument/1147 (accessed 12/04/21).

Ooms, D., Palm, R., Leemans, V. and Destain, M.-F. (2010). A sorting optimization curve with quality and yield requirements. *Pattern Recognition Letters* 31(9), 983–990.

Pronofruit (2021). Pronofruit introduction. Available at: https://pronofrut.cl/en/ (accessed 7/2/2021).

Published by Burleigh Dodds Science Publishing Limited, 2022.

QIMA (2020). Fresh produce quality control – inspections, testing & audits. Available at: https://www.qima.com/testing/produce-inspections-and-quality-control (accessed 20/11/2020).

Rahman, M., Robson, A. and Bristow, M. (2018). Exploring the potential of high resolution WorldView-3 imagery for estimating yield of mango. *Remote Sensing* 10(12), 1866.

SAS (2021). The SurveyMeans procedure: replication methods for variance estimation. Available at: https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=statug&docsetTarget=statug_surveymeans_details51.htm&locale=en (accessed 7/3/2021).

Thompson, A. K. (2003). *Fruit and Vegetables, Harvesting, Handling and Storage* (2nd edn.). Wiley-Blackwell: Oxford, UK. ISBN: 978-1-405-10619-1.

Thompson, S. K. (2012). *Sampling*. Wiley: Germany.

USDA (2021). Grade standards for fruit. Available at: https://www.ams.usda.gov/grades-standards/fruits (accessed 7/3/2021).

Walsh, K. B., McGlone, V. A. and Han, D. (2020). The uses of near infra-red spectroscopy in post-harvest decision support: a review. *Postharvest Biology and Technology* 163, 11140. https://doi.org/10.1016/j.postharvbio.2020.111140.

Wulfsohn, D., Gardi, J., Cohen, O., Garcia-Fiñana, M. and Zamora, I. (2019). Pronofrut: IT-assisted stereology for monitoring orchards for precision horticulture. *Abstract for Presentation at ICSIA 2019 Conference -Mini-Symposium on Sampling in Stereology*.

Zespri (2018). *PG 3 OPC Kiwiflier Spotlight #8*. Zespri OPC, Zespri International Ltd: Mount Maunganui, NZ.